

减少AI应用成本，发挥AI价值

——一个轻量级AI应用实践

北京友谊医院医学数智创新中心 王力华

2025-11

目录

01

背景

02

困境与挑战

03

解决思路

04

轻量级AI应用实践

05

总结



01

背景

2025年南朝HIT论坛课件
版权归演讲者所有

国家战略明确将人工智能列为重点发展领域

人工智能具有明确政策导向，技术发展时机成熟

医疗人工智能进入高质量发展新阶段



中华人民共和国中央人民政府

www.gov.cn

国务院关于深入实施“人工智能+”行动的意见

国发〔2025〕11号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

为深入实施“人工智能+”行动，推动人工智能与经济社会各行业各领域广泛深度融合，重塑人类生产生活范式，促进生产力革命性跃迁和生产关系深层次变革，加快形成人机协同、跨界融合、共创分享的智能经济和智能社会新形态，现提出如下意见。

一、总体要求

以习近平新时代中国特色社会主义思想为指导，完整准确全面贯彻新发展理念，坚持以人民为中心的发展思想，充分发挥我国数据资源丰富、产业体系完备、应用场景广阔等优势，强化前瞻谋划、系统布局、分业施策、开放共享、安全可控，以科技、产业、消费、民生、治理、全球合作等领域为重点，深入实施“人工智能+”行动，涌现一批新基础设施、新技术体系、新产业生态、新就业岗位等，加快培育发展新质生产力，使全体人民共享人工智能发展成果，更好服务中国式现代化建设。

到2027年，率先实现人工智能与6大重点领域广泛深度融合，新一代智能终端、智能体等应用普及率超70%，智能经济核心产业规模快速增长，人工智能在公共治理中的作用明显增强，人工智能开放合作体系不断完善。到2030年，我国人工智能全面赋能高质量发展，新一代智能终端、智能体等应用普及率超90%，智能经济成为我国经济发展的重要增长极，推动技术普惠和成果共享。到2035年，我国全面步入智能经济和智能社会发展新阶段，为基本实现社会主义现代化提供有力支撑。

国务院关于深入实施“人工智能+”行动的意见

人工智能的核心医疗应用场景



中华人民共和国国家卫生健康委员会

National Health Commission of the People's Republic of China

一 人工智能+医疗服务管理

(一) 人工智能+医疗服务

辅 助 学 影 断 像 智 能	智 能 学 辅 助 影 像 质 数 控 据	辅 临 助 床 决 专 策 病 智 能	智 基 能 层 辅 全 助 科 决 医 策 生	辅 医 助 学 治 影 疗 像 智 能	规 划 手 术 智 能 辅 助	智 放 能 射 辅 治 助 疗 勾 靶 画 区	智 能 门 诊 分 诊
咨 智 询 能 就 医	智 能 预 问 诊	智 能 陪 诊	智 能 随 访	调 智 查 能 满 意 度	院 智 后 能 管 患 理 者	辅 智 助 能 生 病 成 历	

卫生健康行业人工智能应用场景参考指引

近千家医院落地大模型本地部署，探索医疗 AI 应用

医院自主部署与应用情况

部署医院地区分布

医院地区	部署医院数量	医院地区	部署医院数量
四川省	56	河南省	19
山东省	50	上海市	17
广西壮族自治区	50	辽宁省	16
广东省	48	新疆维吾尔自治区	15
安徽省	40	宁夏回族自治区	13
江苏省	39	山西省	12
内蒙古自治区	38	贵州省	12
福建省	37	云南省	11
浙江省	33	吉林省	10
湖北省	29	黑龙江省	9
陕西省	27	甘肃省	9
河北省	27	天津市	4
江西省	23	青海省	4
重庆市	22	西藏自治区	0
湖南省	22	海南省	0
北京市	21		

场景	出现次数	场景	出现次数
报告解读	108	风险预警	25
智能导诊	99	智能客服	22
辅助诊断	98	影像分析	22
病历质控	82	科研支持	19
中医相关	69	资源优化	9
病历生成	54	健康宣教	8
智能问答	52	设备调度	8
健康管理	45	运维管理	7
多模态数据融合	39	处方审核	7
临床决策支持	27	合理用药	6
智能诊疗	26	医保控费	2



02

困境与挑战

2025年南网湖上IT论坛课件
版权所有

人工智能落地医院的困境与挑战

1

算法：模型强不等于
适配强

2

算力：投入资金大
高性能 ≠ 高效能

3

医疗数据安全
孤岛效应

4

AI临床应用
普遍“叫好不叫座”

人工智能落地面临门槛高、成本高、预期高的“三高困境”



03

解决思路

2025年南朝HIT论坛课件
版权所有

以轻补强、算力分层、错峰调度，破局“三高困境”



轻量级大模型
(32B千问3)

定义：聚焦特定领域、体量小、部署灵活的AI方案。

核心特点：**低配置**、**高针对性**、**研发成本低**、**快速落地**

作为场景层：**医疗专业能力**，针对不同业务场景微调垂直应用、准确率高的生成、质控类应用



满血版大模型
(671BDeepSeek)

定义：具备强泛化能力、多任务处理能力的大型AI模型。

核心特点：**高算力**、**强通用性**、**开发周期长**、**精度高**

作为基础层：通用能力，复杂推理任务、及时性要求不高的应用、多模态数据分析



轻量级与满血版大模型成本对比

项目	千问3 (32B)	DeepSeek (671B)
GPU数量	1 张	16-32 张
部署成本	约 22万元	350-700万元
电力成本	280元/月	4480- 8960元/月
算法生态	开源模型部署	开源模型部署
响应速度	快	稍慢（依任务、配置而异）
处理复杂任务能力	低	高
成本效益	高	低

合理动态调配算力资源

医院部分AI任务对实时性要求不高，可采用 错峰生成 + 异步处理 + 分层调度

错峰生成 + 异步处理：采用错峰与异步机制，实现非实时任务后台执行，提高算力利用率



白天：算力优先确保关键业务流程不中断



夜间：批量进行非实时推理任务

分层调度：常规任务采用 32B 模型运行，重负载任务再调用满血版算力

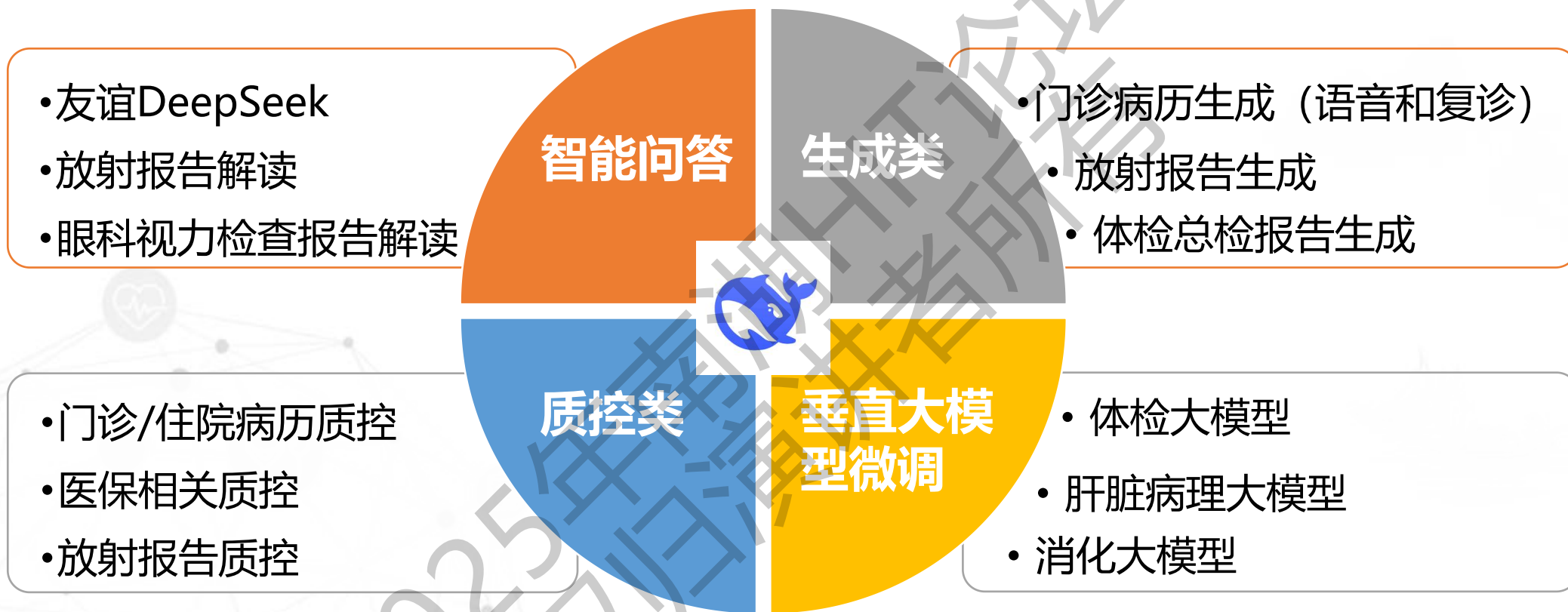


04

轻量级AI应用实践

2025年南粤IT论坛课件
版权归演讲者所有

我院大模型应用场景



轻量级大模型的内涵质控

需求强烈:多科室均有强烈需求,
部分需求实时性要求不高

规则清晰:规范明确、标准统一,
支撑内涵质控高效实施

易见成效:轻量模型适合
快速上手与验证效果

节约成本:选择开源模
型降低开发成本

数据安全:在本地
部署,数据不出院

为什么选这个场景?

轻量化AI质控建设路径

启动前:

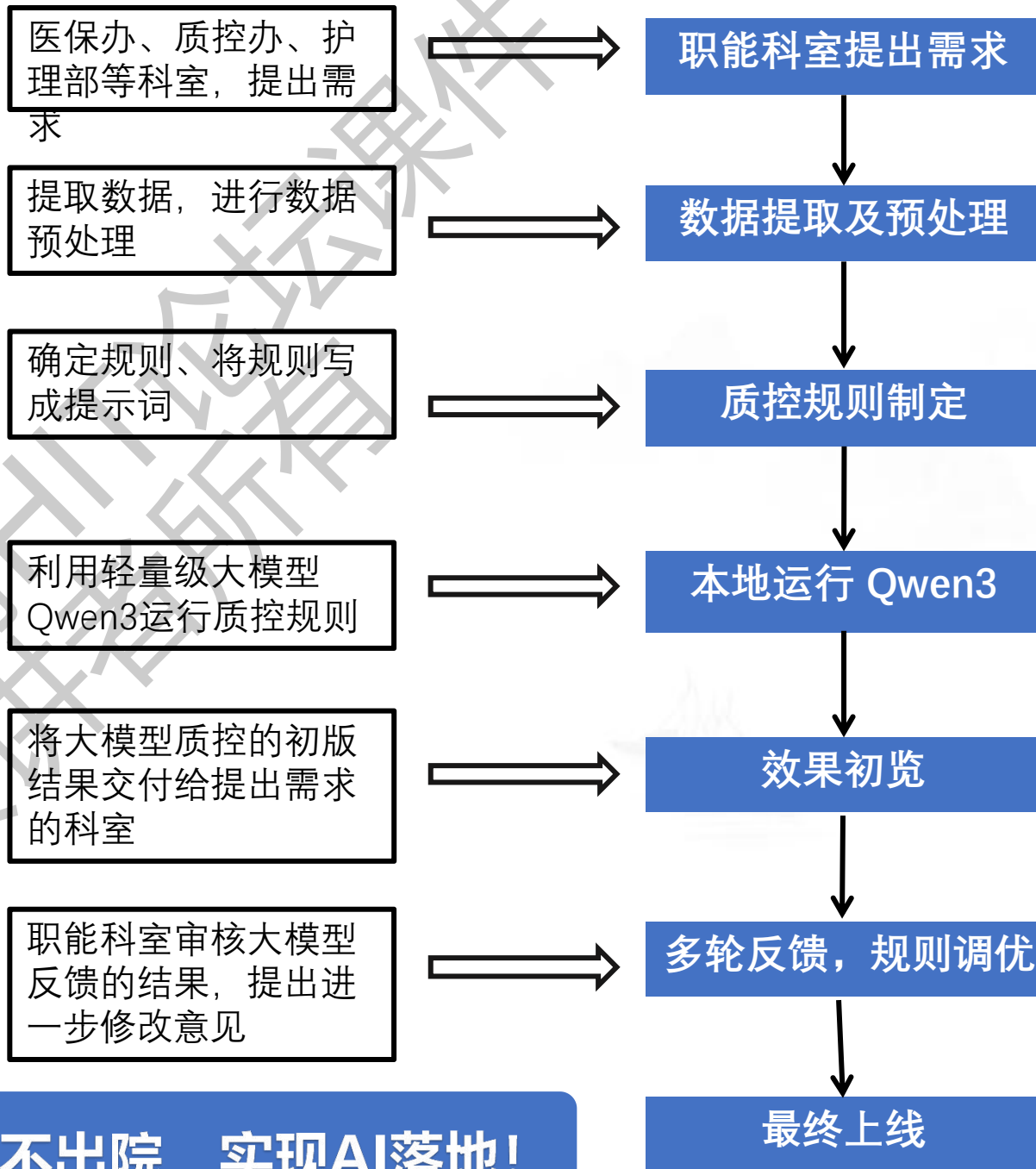
- 讨论确认是否适用于大模型质控工作

过程中:

- 持续优化规则、提高质控覆盖率
- 细化基于风险的质控规则，**专人纵向跟进负责专业组项目质量**

完成后:

将制定的规则**存入本地知识库**中。以便自动化调用，未来通过**智能体**实现按照科室所需频次定期将质控结果返回需求科室，或在信息系统中呈现



需求与数据提取流程



CTLOC_DESC	REA_DESC	院区	医保类型	CHIEF_COMPLAIN	HISTORY_ILLNESS	DISCHARGE_DIAGNOSIS
脊柱外科	基本医疗保险	通州院区	医保	腰痛1天^低危	患者自述约1天前不慎	脊柱骨折^腰痛^高
通州肝病二科(重症肝病)	异地持卡	通州院区	医保	双下肢水肿伴乏力	患者2月余前因药物	
通州血液内科	异地持卡	通州院区	医保	反复发热3年余	3年余前(2022.1月	
关节外科	基本医疗保险	通州院区	医保	右侧肩关节疼痛伴	患者于7月前无明显	
风湿内科	基本医疗保险	通州院区	医保	间断关节肌肉疼痛	1患者10余年前无明	
通州血液内科	异地持卡	通州院区	医保	间断发热2年余。	患者2年余前无明显	
肝病内科	基本医疗保险	通州院区	医保	发现肝功能异常6天	患者6天前无明显诱	
普外分中心	异地持卡	通州院区	医保	头晕、嗜睡、皮肤	患儿20天前无明显	
骨科	异地持卡	西城	未结算	自行不慎摔伤一日	患者今日晨不慎摔	
心脏中心	基本医疗保险	西城	医保	血压控制不稳10余	患者既往高血压2	
肿瘤科	异地持卡	西城	医保	确诊肺恶性肿瘤1年	患者1年无明显诱因	肺恶性肿瘤(T1N

入院记录伤因的内涵质控

- 1、针对“入院记录”全部内容进行内涵质控，包括主诉、现病史、既往史、个人史、入院诊断等。
- 2、患者的“身份类别”：对持卡实时结算 医保类险种的进行核查，包括：基本医疗保险、军休公费医疗、离休统筹、超转人员、公疗医照、无保障老年人、无业居民、学生儿童、异地持卡。

制定质控规则

国家卫生健康委员会办公厅

国卫办医政函〔2025〕227号

国家卫生健康委办公厅关于印发 医疗质量安全核心制度落实情况 监测指标(2025年版)的通知

指标十七、术前讨论计划手术一致率

定义：实际开展手术与术前讨论计划手术一致的手术例

数占同期手术总例数的比例。

根据规则制作提示词

SYSTEM_PROMPT = ""你是一个医疗质量控制的专家，负责审核术前讨论与手术记录中的手术名称是否匹配。你的任务是对比“术前讨论中的手术名称”和“手术记录中的手术名称”，判断两者是否指的是同一种手术。

核心指令

1. 匹配标准：不要求文字完全一致，只要表达的是同一种手术（包括同一手术的不同名称或简称）即视为匹配。
2. 注意事项：
 - 需考虑手术的主要操作和部位是否一致
 - 需考虑是否为同一疾病的同一治疗方式
 - 需注意区分名称相似但实际不同的手术

术式匹配细化规则

根据医院各科室书写习惯制定特殊规则

1. 首先检查患者住院信息：
 - 如果患者来自日间病房，直接标记为“不适用”（不参与匹配判断）
 - 否则继续执行后续匹配规则
2. 特殊匹配规则（以下情况应视为匹配）：
 - (1) 当术前术式为“宫腔镜检查”或“宫腔镜检查，备一期手术”或“宫腔镜手术” → 手术记录包含“宫腔镜”即视为匹配
 - (2) 当术前术式为“鼻内镜下鼻腔鼻窦手术” → 手术记录包含“鼻内镜”即视为匹配
 - (3) 当术前术式为“痔切除术” → 手术记录包含“内痔套扎术”或“痔切除术”即视为匹配
 - (4) 当术前术式为“腹腔镜下胆囊切除术”或“腹腔镜胆囊切除术” → 手术记录包含“腹腔镜”和“胆囊切除”即视为匹配

质控多轮调优

手术记录	手术记录中的手术名称	AI质控结果	匹配说明
footer:	经尿道输尿管镜激光碎石术+输尿管支架置换术	不匹配	术前小结总结的术式为“经尿道输尿管镜激光碎石术”，而手术记录中的手术名称为“经尿道输尿管镜激光碎石术+输尿管支架置换术”。
footer:	经肛门内痔套扎术+外痔切除术+混合痔注射+肛管	不匹配	术前小结总结的术式为“痔切除术”，而手术记录中的手术名称为“经肛门内痔套扎术+外痔切除术+混合痔注射+肛管肿物切除术”。虽然都
footer:	经肛门内痔套扎术+外痔切除术+混合痔注射	不匹配	术前小结总结的术式为“痔切除术”，而手术记录中的手术名称为“经肛门内痔套扎术+外痔切除术+混合痔注射”。虽然两者都涉及痔疮的
footer:	宫腔镜下宫颈息肉电切术+诊刮术	不匹配	术前小结总结的术式为“宫腔镜检查，备一期手术”，主要描述的是一个初步的宫腔镜检查，并准备进行进一步的一期手术。而手术记录中的
footer:	鼻内镜下左侧上颌窦开放	不匹配	术前小结总结的术式是“鼻内镜下鼻腔鼻窦手术”，这是一个较为广泛的描述，可能包含多种具体的鼻窦手术。而手术记录中的手术名称是“
footer:	经肛门内痔套扎术+外痔切除术+混合痔注射	不匹配	术前小结总结的术式为“痔切除术”，而手术记录中的手术名称为“经肛门内痔套扎术+外痔切除术+混合痔注射”。虽然两者都涉及痔疮的
footer:	经尿道输尿管支架置入术	不匹配	术前小结总结的术式是“经尿道输尿管镜激光碎石术”，而手术记录中的手术名称是“经尿道输尿管支架置入术”。前者是通过输尿管镜使用
footer:	经肛门内痔套扎术+外痔切除术+混合痔注射+肛	不匹配	术前小结总结的术式为“痔切除术”，而手术记录中的手术名称为“经肛门内痔套扎术+外痔切除术+混合痔注射+肛裂切除术”。虽然都涉及
footer:	宫腔镜下子宫内息肉电切术+诊刮术	不匹配	术前小结总结的术式为“宫腔镜检查，备一期手术”，这表明术前计划主要是进行宫腔镜检查，并视情况决定是否进行进一步的手术。而手术

术式匹配细化规则

- 1、添加入院时间和出院时间或者患者的住院科室，去除日间病房的患者。
- 2、不匹配的数据里面，
 - (1) 如术前小结中总结的术式为：宫腔镜检查，备一期手术，要求大模型去比对手术记录中的手术名称是否宫腔镜相关，相关即认为匹配。
 - (2) 如术前讨论总结的术式中为：鼻内镜下鼻腔鼻窦手术，要求大模型去比对手术记录中的手术名称是否为鼻内镜下的相关手术，相关即认为匹配。
 - (3) 如术前小结中总结的术式为：痔切除术，要求大模型去比对手术记录中的手术名称是否含有：内痔套扎术，含有即认为匹配。

数智中心

提出修改意见

交付大模型质控结果

职能科室

智能质控效果展示（应用CDSS或病历形式质控消息窗口）

日期: 2019-09-25 11:44
今日查房, 患者无明显不适主诉, 查体: 发育正常, 营养良好, 大便正常, 小便解出困难, 查体: 神清, 精神可, 皮肤, 巩膜无黄染, 全身浅表淋巴结未及肿大。头部可见外伤所致片状淤青, 头颅大小正常, 无畸形, 双侧瞳孔等大等圆, 对光反射存在, 颈软, 气管居中, 颈静脉无怒张, 甲状腺不大, 胸廓对称, 双肺呼吸音清, 未闻及干、湿罗音, 心率72次/分, 律齐, 各瓣膜听诊区未闻及杂音, 脊柱, 四肢未见畸形, 生理反射存在, 病理反射未引出, 双侧瞳孔大小不等, 无明显压痛, 活动度可。辅助检查: 血常规分析(五分类仪器检测法): 白细胞计数 $4.94 \times 10^9/L$, 红细胞计数 $3.57 \times 10^{12}/L$, 血红蛋白 $99g/L$, 血小板计数 $157 \times 10^9/L$, 中性粒细胞比率 47.04% , 溶血功能检查: 凝血酶原时间国际标准化比值 1.24 , 纤维蛋白原 $4.92g/L$, 抗凝血酶 53% , 超敏C反应蛋白测定(速率散射比浊法): 超敏C反应蛋白 $60.5mg/L$, 病毒八项: 乙肝核心抗体(定量) $8.01s/co$, 生化全套+电解质: 钾 $3.45mmol/L$, 钙 $2.59mmol/L$, 白蛋白 $30.7g/L$, 直接胆红素 $8.0umol/L$, 乳酸脱氢酶 $397U/L$, 组胺免疫球蛋白测定+血清标本 β_2 微球蛋白测定(化学发光法): 补体C3 $0.592g/L$, 超敏C反应蛋白 $58.8mg/L$, β_2 微球蛋白 $0.93mg/L$, 钾钙磷原正常, 张李平副主任医师查房: 患者入院后测得体温 $38^\circ C$, 予百美诺静滴后体温恢复正常, 关注患者体温, 必要时加用抗生素, 诉排尿困难, 查体示膀胱充盈, 嘱...

临时医嘱 患者信息: 0000342241 *** 男 94 血液科一 病历评分:92 甲级病历 责任医师: 血液科一/张李平(00...)

日期	时间	医嘱内容	医师签名	执行时间	执行人签名	医嘱执行状态	医嘱状态
2019-09-24	21:10	静脉输液每瓶加收立即使用	张李平	2019-09-24 21:13	陈南清	已执行	执行
2019-09-24	21:10	盐酸莫西沙星注射液(佰美诺) 静滴 12 立即使用	张李平	2019-09-24 21:13	陈南清	已执行	执行
2019-09-25	08:40	CT盆腔常规增强扫描立即使用	张李平	2019-09-25 10:33	毛春花	已执行	执行

质控规则: 0323342
问题: 病程中未在3天内体现医嘱中抗生素使用, 医嘱日期:2019-09-24 21:10:28, 医内容:盐酸莫西沙星注射液(佰美诺)

病程中未在3天内体现医嘱中抗生素使用, 医嘱日期:2019-09-24 21:10:28, 医内容:盐酸莫西沙星注射液(佰美诺)

- 医师书写病历进行内涵质控（质控当前、质控全部）
- 系统提示质控消息，方便医师进行修改
- 根据规则进行自主逻辑判断，提高医生书写病历的准确性

例：入院记录中未在3天内体现医嘱中抗生素使用

实践成效与价值

与相关科室紧密合作，进行迭代调优

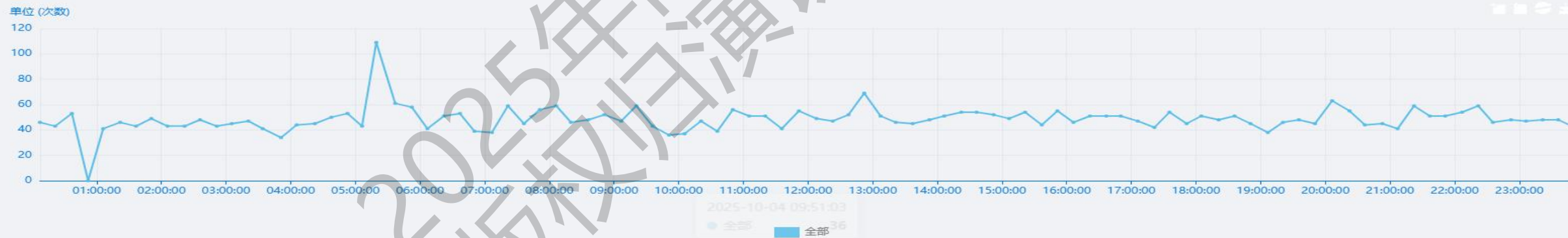
- 在开展内涵质控工作以来，已完成包括病历内涵质控、危急值质控、二线三线查房质控、伤因分析质控、特殊使用级抗生素质控、输血护理记录合规性质控等**在院患者**质控任务。
- 总质控病例数：**32762份**
- 总质控迭代轮次：**36次**
- 准确率：提升至**95%**

算力使用分析

开展质控工作前接口调用情况：

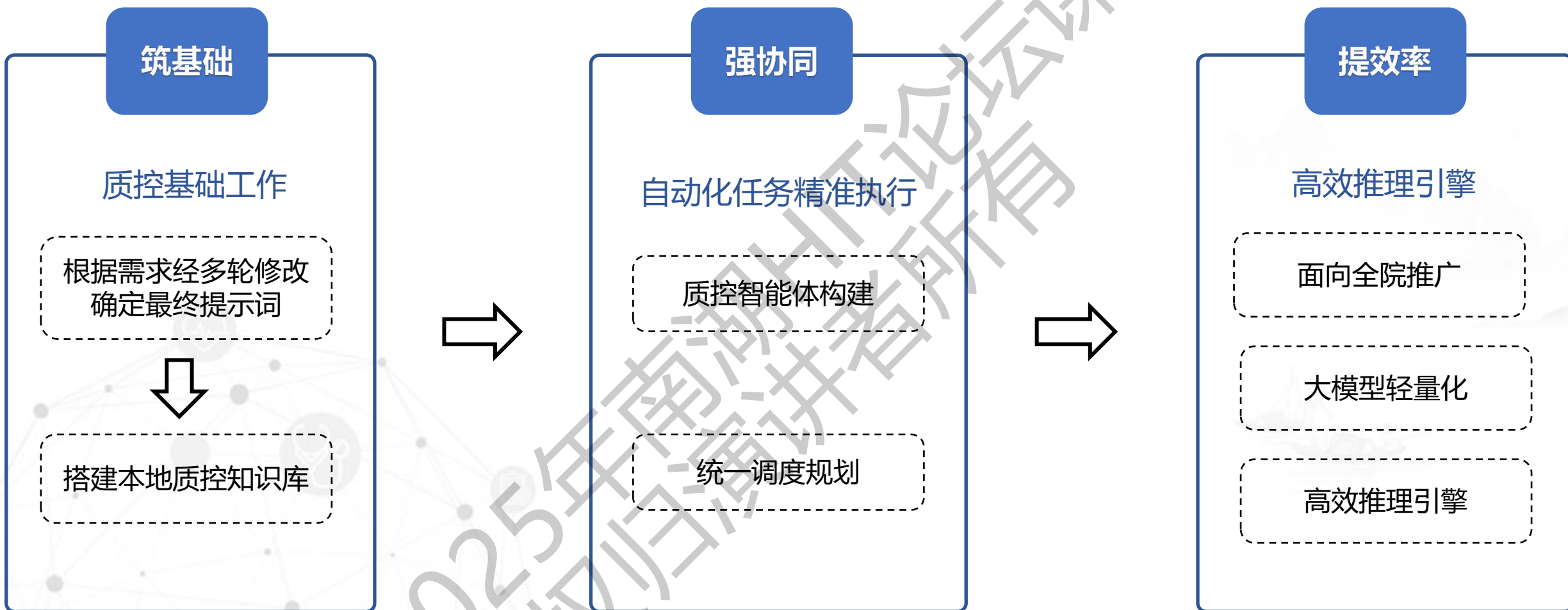


开展质控工作后接口调用情况：



提高算力使用率，减少不必要资源消耗，以更低的成本发挥 AI 价值

赋能全流程内涵质控





05

总结分析

2025年南朝HIT论坛课件
版权所有

轻量级 AI 质控：实现低成本高可用

核心理念

- ◆ **轻量级 AI**：基于开源模型、低研发成本，结合规则优化减轻推理难度，单卡即可快速部署医院并完成非实时任务，提升诊疗效率；

特征亮点

- ◆ **低成本**：单卡运行，硬件与运维投入最小化
- ◆ **规则驱动**：以明确的质控规则减轻推理负担
- ◆ **错峰异步**：非实时任务后台执行，提高算力利用
- ◆ **低门槛**：基于开源模型即可落地，无需自研

轻量级 AI 的可扩展场景：

医保质控

- 对接医保办规则库，实现实时合规性审核与违规项预警

护理文书

- 自动检测护理记录规范性与完整性，减轻护士书写负担

患者召回追踪

- 通过对患者检查检验指标进行综合性分析，同时对高风险患者进行召回

其他潜在场景

- 影像筛查报告自动生成、手术记录合规性审核、门急诊病历逻辑一致性检查等

精准匹配需求，实现成本与价值平衡



轻量级大模型：“小而精”

适合低成本、低研发、快落地的需求场景。



大模型：“大而全”

适合高需求、长周期、强综合能力的场景。

算法务实 —— 从业务场景出发，强调可落地与实用价值。

算力高效 —— 通过错峰运行与轻量级推理，实现成本与性能的平衡。

数据成长 —— 建立持续反馈机制，推动模型在临床场景中不断优化。

感谢聆听!